DOCUMENT RESUME

ED 205 537 TM 810 410

AUTHOR Smith, Richard M.
TITLE The Assessment of Partial Knowledge: The Use of the Credit Model in Scoring Multiple-Choice Items.
PUB DATE Mar 81
NOTE 34p.: Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Algebra: High Schools: Latent Trait Theory: *Multiple Choice Tests: *Scoring: *Test Items: *Weighted Scores
IDENTIFIERS *Distractors (Tests): New Jersey College Basic Skills Placement Test: *Partial Credit Model: Rasch Model

ABSTRACT

One of the recurrent themes of the psychometric literature has been the idea that the incorrect responses a person makes to test items contain information that might be useful in determining the person's position on the variable the items are intended to define. The "Partial Credit" model, a member of the family of latent trait models proposed by Rasch, provides a means of assessing this partial knowledge while maintaining the desirable psychometric characteristics that all Rasch models enjoy. In this study the results of 30-item elementary algebra test are analyzed using the standard dichotomous scoring procedure and several procedures for assessing partial knowledge. The preliminary results indicate that the a priori ordering of distractors using integer weights provides useful information, particularly in the lower ability ranges. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED205537

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

THE ASSESSMENT OF PARTIAL KNOWLEDGE: THE USE OF THE CREDIT MODEL IN SCORING MULTIPLE-CHOICE ITEMS

RICHARD M. SMITH

UNIVERSITY OF CHICAGO

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. M. Smith

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

March 1981

TM 810 410

Paper presented at the annual meeting of the American Educational Research Association
Los Angeles, April 1981

ABSTRACT

One of the recurrent themes in the psychometric literature has been the idea that the incorrect responses a person makes to test items contain information that might be useful in determining the person's position on the variable the items are intended to define. The "Partial Credit" model, a member of the family of latent trait models proposed by Rasch (1960), provides a means of assessing this partial knowledge while maintaining the desirable psychometric characteristics that all Rasch models enjoy. In this study the results of a thirty item elementary algebra test are analyzed using the standard dichotomous scoring procedure and several procedures for assessing partial knowledge. The preliminary results indicate that the a priori ordering of distractors using integer weights provides useful information, particularly in the lower ability ranges.

THE ASSESSMENT OF PARTIAL KNOWLEDGE: THE USE OF THE CREDIT MODEL IN SCORING MULTIPLE CHOICE ITEMS

INTRODUCTION

One of the recurrent themes in the psychometric literature is the idea that the incorrect responses an individual makes to test items contain information that might be useful in estimating the person's position on the variable that the test is designed to measure. The usual approach, particularly in aptitude and achievement measurement, has been to score items dichotomously, treating one response as correct and all other responses as equally incorrect. This system of scoring has been perpetuated by the wide-spread use of multiple choice items and computer scoring. The typical scoring procedure awards one point for a correct response and zero points for an incorrect or omitted response. A correction may also be made for guessing in this scoring procedure. In most cases this correction is made by subtracting a fraction of the number of incorrect responses, omits excluded, from the total number of correct responses. But this scoring system ignores the possibility that incorrect responses may contain information about individual differences in achievement or apitude.

Consider the following multiple choice item as an example:

> As opposed to measurement and evaluation, assessment is more likely to describe individuals in
>
> A) largely qualitative terms.
> B) an equal combination of qualitative and quantitative terms.
> C) quantitative terms with respect to social values.
> D) largely quantitative terms.

If alternative A is the most correct response, under the dichotomous scoring system anyone choosing A will receive one point and anyone choosing B, C, or D will receive 0 or -1/4 points, depending on whether or not a guessing correction is used.

If the purpose of administering this item is to quantify an individual's knowledge of assessment and then to use this information to make recommendations about possible instructional strategies which might further increase this knowledge, it may be useful to examine the incorrect alternatives to assess the level of knowledge represented by each. It could be argued that individuals choosing response B have some knowledge of what assessment is. They do not understand the distinction between measurement, evaluation, and assessment as well as someone who chooses alternative A, but they certainly have a better understanding than someone who chooses C or D. If this is the case, it seems unfair to give a person choosing alternative B the same point value as a person choosing response C or D. We may obtain a better picture of a person's ability if we award 2 points for alternative A, 1 point for alternative B, and 0

points for alternatives C and D.

This type of argument has been the major motivational factor in the search for a scoring procedure which utilizes the information contained in incorrect responses. These procedures have been described as multiple category scoring, weighted response scoring and partial knowledge scoring, to name just a few. This search has resulted in an interesting variety of scoring methods as well as a variety of novel item types.

REVIEW OF SCORING PROCEDURES

The recent efforts to measure partial knowledge tend to take one of four basic forms; item weighting, response weighting, elimination scoring, and degree of confidence scoring. The item weighting methods (Kelly 1934, Stanlaker 1938, Guilford 1941, Cronbach 1941) involve the weighting of only the correct responses. In this method the items are scored in the conventional dichotomous scheme, 1 point for a correct response and 0 points for all other responses. The item weights are applied after scoring, consequently the weights are applied only to items that the person answers correctly. These weights may be based on a variety of factors ranging from the judged importance of the item to the point biserial correlation between the item and the total test score over some "standard" sample.

Response weighting methods (Smith & Tyler 1942, Nedelsky 1954, Ramsey 1968, Hendrickson 1971, Patnaik & Traub 1973, Reilly

& Jackson 1973, Echternacht 1976, Waters 1976, Bejar & Weiss 1977, Claudy 1973) involve the weighting of all responses in a multiple choice item. These weights serve to differentiate between the various incorrect responses. This method is based on the assumption that the incorrect responses represent a useful source of variation which can be added to the variation introduced by the differences between the correct and incorrect responses to increase the utility of the total score. There are two basic procedures for deriving response weights, the a priori method which involves estimating the theoretical degree of correctness for each response, and the empirical method which involves utilizing data from a "standard" test administration to produce the individual response weights.

Elimination scoring (Coombs 1953, Coombs, Milholland, & Womer 1954, Davis 1959, Davis & Fifer 1959, Cross, Thayer, & Frary 1980) requires the examinee to indicate the incorrect responses to an item rather than to select a correct response. This method is based on the assumption that there are three levels of knowledge, complete knowledge , i.e. being able to eliminate all of the incorrect responses, partial knowledge, i.e. being able to eliminate some of the incorrect responses, and misinformation, i.e. believing that the correct response is incorrect. Under this scoring method it is common to award 1 point for each incorrect and to penalize the examinee 3 points if the correct response is eliminated. The scoring rules vary, however, depending on the author and the number of responses in a

given item.

Degree of confidence scoring (Dressel & Schmid 1953, Hambleton, Roberts, & Traub 1970, Diamond 1975, Poizner, Nicewandner, & Gettys 1978) is based on the assumption that examinees possessing partial knowledge can eliminate some of the incorrect responses but can not differentiate among the remaining responses without guessing. There are several methods of utilizing this concept in scoring tests. They range from corrections for gross errors to ordinal ranking of responses by degree of correctness to a probabilistic weighting of all responses according to degree of confidence in that response as the correct answer.

Most of the studies reviewed here use the classical true-score psychometric model as the framework for developing the scoring procedure. The recent development of latent trait models, in particular the Rasch one item parameter model and the Samejima two item parameter multiple category model (Thissen 1976), has led to a series of methods based on one of the latent trait model rather than the true-score model.

In the classical true-score model there are two indices for judging the utility of a scoring model, the reliability of the test and the validity of the test. In most instances the test reliability was calculated using the standard measures of reliability, KR-20, Cronbach's alpha, split-half, or parallel forms. Test validity, when reported, was estimated by using concurrent validity, calculcating the correlation of the test

score with another criterion or test score for each person . This score was obtained at some time near the administration of the experimental form. The studies which utilize the logistic model have used, in addition, the concept of test information over the range of abilities measured by the test as a means of determining if the additional variation introduced by the assessment of partial knowledge is related to ability or if it is some form of error variation.

THE STUDY

In exploring the possible methods of weighting responses for use with the Partial Credit model and the effect of these different methods on item difficulties and person abilities we have examined an elementary algebra test, a subsection of the New Jersey College Basic Skills Test. This test was administered to 47,000 students entering public colleges in New Jersey between May 1979 and September 1979. This was a fairly difficult test with the state-wide mean at 13.8 ,of a possible 30, items correct. The test was composed of items from two basic content areas; solving equations and algebraic operations.

The purpose of this study is to investigate the applicability of the Partial Credit model to multiple choice items with a hierarchy of responses. The hierarchy referred to here is the capacity to order the responses to a given item from the least correct to the most correct. The underlying construct, discussed earlier, is that even the incorrect responses to

multiple choice test items contain information about the ability of the individual being measured. The choice of one incorrect response may indicate a higher level of understanding or ability than the choice of another incorrect response.

This is certainly true in mathematics where a single question may require the successful completion of several steps to produce the correct result. In mathematics, an incorrect answer resulting from the misapplication of a mathematical theory or principle would indicate a lower level of mathematical ability than an incorrect answer which resulted from the correct application of the theory or principle with a computational error. Consider the problem:

$\sqrt{12} + \sqrt{27} =$

A) $3\sqrt{5}$
B) $\sqrt{39}$
C) $5\sqrt{3}$
D) $3\sqrt{2} + 3\sqrt{3}$

Alternative C is the correct response and should receive full credit. However a person choosing response D exhibits a higher level of mathematical ability than a person choosing response B. In response D the ability to reduce $\sqrt{27}$ to $3\sqrt{3}$ indicates a rudimentary knowledge of radicals and interchanging the 2 and the 3 in reducing $\sqrt{12}$ could easily be a clerical error. The simple summing of the numbers under the radical sign, necessary to arrive at $\sqrt{39}$ in choice B, indicates a lack of even rudimentary knowledge of radicals. Response A is somewhat more difficult to classify. This type of response could possibly come

from properly simplifying $\sqrt{27}$ to $3\sqrt{3}$, incorrectly simplifying $\sqrt{12}$, then improperly summing the terms under the radicals after having removed the common factor of 2

$$3\sqrt{2} + 3\sqrt{3} = 3(\sqrt{2} + \sqrt{3}) = 3\sqrt{5}$$

or from simply transposing the 3 and 5 in the final step of the correct solution. This error would seem to imply a better understanding of radicals than the similar type of error involved in choosing response B, but this is a subjective judgement. It is difficult to imagine that a person who could simplify one of radicals and remove the common factor would then simply sum the two terms under the radical. The proper solution to this classification dilemma would be to replace this response with one that contained a more plausible type of error that would easily fit into the hierarchy of the other three responses.

If an objective hierarchial ordering of the responses to a given item is possible, the next consideration is the assignment of weights to each of the possible responses. A review of the numerous studies utilizing weighted responses (Hambleton, Roberts, & Traub 1970, Hendrickson 1971, Reilly & Jackson 1973, Patnaik & Traub 1973, Echternacht 1976, Bejar & Weiss 1977) indicates that there has been considerable variation in the nature of the response weighting schemes employed. One source of disagreement in these studies is practical, while the other disagreement appears to be philosophical in nature. We will address the philosophical disagreement first.

The philosophical issue is the one raised by the comparison

of a priori response weighting and empirical response weighting and it bears careful attention. The issue is one that has its roots in one of the fundamental questions of science: should a theory be developed on the basis of theoretical concerns which must be confirmed or refuted by empirical evidence, or should a theory derive its structure from the emiprical data? The use of an empirically derived weighting system seems to have two flaws; the first is that this type of empiricism runs contrary to the accepted practices of science, and second, this practice allows the person constructing the items and ultimately the variable which will be defined by these items, to proceed without committing himself to a theoretical definition of the variable that must be confirmed or refuted by the data.

The practical question concerns the sample dependent nature of these weights. Within the context of measurement, in particular the Rasch measurement model, there is an additional problem, namely the sample dependency of the empirical response weights. The use of sample dependent response weights in this situation seems to be anomalous. Is the scoring procedure to be redefined each time a test is administered, with the weights each time depending on the sample, or is a constant set of response weights to be employed with the ensuing requirement of equivalent samples necessary for useful results? In either case the use of the empirical weighting procedure reduces the generalizability of the weights and, therefore, limits the applicability of the technique even with constant re-evaluation.

Finally there is a third problem which is not discussed in these studies, it is the problem of omitted respnses. Earlier we noted the inconsistency of the results reported in studies comparing empirical and a priori weighting systems. In these studies there would appear to be a systematic relationship betwen these results and the presence of omits. The Bejar and Weiss (1978) study used simulated results with no omits and found a priori scoring better than empirical scoring. Hambelton, Roberts, and Traub (1970) used fairly easy tests with the consequent relative scarcity of omits in their study and found that the two methods, a proiri and empirical, produced equal measues of reliability for the test under study. Echternacht (1975) and Reilly and Jackson (1973) used fairly difficult and speeded tests, resulting in more omits, and found the empirical weighting system to be better. Omits, as the analysis reported in this paper will show are a sample dependent phenomena. The issue of how to effectively deal with omits has not only been hampered by the sample dependent nature of the omitting behavior, but it has also been hampered by the nature of classical measurement theory which requires that all persons receive the same set of items if the scores are to be comparable.

If we look more closely at omitting behavior we can see that it can come from one of two sources: either the person never attempted the item, or the person read the item and decided, for whatever reason, not to answer. It is impossible to distinguish between these two possible sources in scoring a multiple-choice

test. If what we are truly interested in is the interaction between the person and the item, then it doesn't make sense to lump these two distinct behaviors together and to include both of them in the ability estimation process, since one does not involve any interaction between the person and the item. Nor does it make any sense to attempt to construct an artifical rule to distinguish between the two behaviors in order to eliminate one of them from ability estimation. Since all Rasch measurement models make possible item-free measurement it would seem most appropriate to simply exclude all omitted items from the process and base the person's ability solely on the items that we are certain that the person interacted with in the prescribed fashion, thus avoiding the problem of omits entirely.

ESTABLISHING ITEM RESPONSE WEIGHTS

These problems are illustrated by the following example:

> If the sum of three numbers is 90 and one of the numbers is x,what is the sum of the other two?
>
> A) 90 + x
> B) 90 + 2x
> C) 90 - x
> D) 90 - 2x

There are two decisions to be made in answering this item. First, should the unknown be of the form "x" or "2x", and second, should the operation used to express the "sum of the other two" be addition or subtraction? In considering the mathematical

correctness of the various responses, it was the subjective opinion of the author that the identification of the proper form of the unknown, "x" in this case, was indicative of a higher level of algebraic ability than the choice of the proper operation. The following a priori weights were then assigned:

Choice	Correctness	Weight
A. 90+x	incorrect operation correct unknown	2
B. 90+2x	incorrect operation incorrect unknown	0
C. 90-x	correct operation correct unknown	3
D. 90-2x	correct operation incorrect unknown	1

Getting both parts right yielded the maximum score, getting both parts wrong yielded the minimum score, and getting one part right yielded intermediate scores with the advantage going to the person demonstrating the more advanced skill, the choice of the proper unknown.

If we were to employ empirical weighting for this item based on the number of people responding to each item, we would employ the following weights:

Choice	Number	Weight with omit excluded	Weight with omit included
A. 90+x	140	0	1
B. 90+2x	362	1	2
C. 90-x	1548	3	4
D. 90-2x	537	2	3
Omit *	76	-	0

* Based on a sample of 3151. 341 of the original 3151 people were

not considered due to the fact that they did not answer 3 items on the test correctly. People with a raw score of 3 or less were not considered to have "taken the test".

In this weighting scheme persons with the least theoretically correct response, B, would receive a higher score than those persons who chose the more mathematically correct response, A. Persons who knew that you have to subtract an unknown from the total to get one of the parts, response D, would get two points more than a person who could identify the proper numerical coefficent of the unknown, response A.

If we were to use the average ability level of the individuals choosing a particular response as the criterion for assigning the empirical weights, we would get yet another ordering. In this case we are using "equated deltas" for each response based on a sample of 3151 individuals less the 341 who did not answer three items correctly. The "equated delta" has a mean of 13 and a standard deviation of 4. These weights are used primairly for item analysis and test equating but do represent the average ability level for individuals choosing a given response. In this case the following weights would be assigned:

Choice	Delta	Weight
A. 90+x	11.0	2
B. 90+2x	10.9	1
C. 90-x	14.8	4
D. 90-2x	12.4	3
Omit	9.8	0

In this instance response B, the least correct response, would

receive the lowest weight among the possible responses. The more mathematically correct response, A, would receive a lower weight than the less mathematically correct response, D.

If we look at the 500 individuals chosen for this study, we would find the following frequency of responses:

Choice	Number	Weight with omit excluded	Weight with omit included
A. 90+x	17	0	0
B. 90+2x	61	1	1
C. 90-x	248	3	4
D. 90-2x	86	2	2
Omit	88	-	3

If we choose to exclude the omit from the weighting or consistently weight it zero, we have exactly the same weighting we would have if we employed the frequency weights discussed above, but if we were to include the omit in the frequency weights there would be a considerable reordering of the weights using this sample.

The preceding analysis points out the contradictory patterns which can result from the use of empirical weighting schemes. An important point follows from this. If we are to understand the items and hence the variable, we must have a notion of the correctness of the responses which is derived independently of the data. If we are to generalize to future administrations, we must have a weighting scheme which is freed from the characteristics of the current sample. In essence, it is necessary not only to write the items with the variable in mind,

i.e., constructing the items so as to define clearly a number of steps or positions along the variable, but also to construct the distractors for these items with the same variable in mind, choosing the distractors to represent clearly defined positions on the variable.

THE RESULTS

We have compared four types of scoring procedures for the elementary algebra test; dichotomous (scored 1,0,0,0,0), 5 category a priori (scored 4,3,2,1,0), 4 category a priori (scored 3,2,1,0,0), and 5 category empirical (scored 4,3,2,1,0). In these four scoring methods the responses are ordered from most correct to least correct acording to the particular method. The fifth category is for omits. This comparison was based on a sample of 500 individuals drawn from the state-wide sample. In order to judge the effect of the presence of omits on item difficulty and person ability a second sample of 500 individuals was drawn. In this case individuals with omits in their response strings were deleted. This sample was then used with two scoring procedures; dichotomous (scored 1,0,0,0) and 4 category a priori (scored 3,2,1,0).

The tenative results of these investigations yeilded the following suggestions for the analysis and scoring of multiple choice data.

1. Although item fit and person fit are sample dependent statistics, and one would expect sample-based empirical weighting to produce better fit

to the model, the a priori scoring procedure produced item and person fit statistics which were as good if not better than the empirical weighting. (The complete analysis on which this conclusion is based is presented in Smith (1981).)

2. The awarding of partial credit for incorrect responses introduces an additional source of variance into the ability distribution. This can be seen in two forms; first, the plot of the dichotomous abilities against the abilities obtained from the Credit model scoring indicate a one to one relationship at the higher ability levels, but at lower ability levels (13 correct or less) there is considerable variation in the Credit ability estimates for a given dichotomous ability (see graph 1). The second illustration of this additional variance is in the item plots (see plot 1). Thissen (1976) contends that one indication of information gained from incorrect responses is the fact that different incorrect responses will be more probable at different points in the ability distribution. As we can see in plot 1, where the correct resopnse is labeled 3, the incorrect response weighted 1 is the most probable response for a fairly sizable part of the ability range, thus indicating a gain in information by differentiating between the incorrect responses in this item.

For the practitioner the Credit analysiis offers two distinct advantages over the traditional dichotomous Rasch analysis. First, it can provide an additional source of information for estimating abilities at the lower end of the ability distribution. If one is interested in making decisions in the raw score range running from 60% correct and up, this added variance will not be of great advantage. However, if one is forced to make decisions about people in the dichotomous raw score range running between 30% and 60%, this added variance can be potentially useful. Second, the results of the Credit analysis

can be very useful in revising items. The BICAL output provides very little information about the individual distractors in each item other than the distribution of responses. Similarly, the fit statistics provide information about the item as a whole but provide little information about the individual distractors. The Credit analysis provides individual item plots which show the probability of each response over the entire ability range. This, as we will demonstrate, can be a powerful tool in the item review process.

In plot 1 we see the probablity curves for item 8 on this test. The distribution of responses for this item was:

response weight	0	1	2	3
no. of responses	28	190	29	259

From these curves we can see that the three incorrect responses appealed to three different ability groups. The separation in terms of ability between the worst response (plotted as a 0) and the most correct response (plotted as a 3) is considerable. One response (plotted as a 2) in this question seems to be contributing very little to the functioning of the item. This response is never very probable. In attempting to make this a better item we may wish to seek a distractor that would have a stronger appeal in the 0-1 logit range.

In plot 2 we see the probability curves for item 2 on this test. The distribution of responses for this item was:

response weight	0	1	2	3

no. of responses	254	11	12	266

As we can see from the distribution of responses and the probability curves, this item is only running on two cylinders. Two of the responses are almost useless. The probability of responding to the responses weighted 1 and 2 is so small that they have a negligible impact on the probability of the other two responses. Despite having four choices this item might just as well have two choices (0 and 3) and be scored dichotomously. In order to correct this situation two more plausible distractors must be found.

In plot 3 we see the probability curves for item 12 on this test. The distribution of responses for this item was:

response weight	0	1	2	3
no. of responses	33	91	154	225

In looking at the probability plots for this item we can see that both the response with the weight of 1 and the response with the weight of 2 have areas in the ability distribution where they are the most probable response. One indication of this is the ascending order of the step values. In the case where one or more distractors is not working the step values would not be in ascending order as in item 8 and 2. This item comes as close to the ideal item as any on this test. The ability range between where the response weighted 0 and the response weighted 3 is the most probable is a little small, approximately .5 logits. The item sould function better if these two curves were a little

further apart.

To demonstrate that the frequency of responses does not provide exactly the same information as the probability plots, we consider item 27 shown in plot 4. The distribution of responses for this item was:

response weight	0	1	2	3
no. of responses	122	82	120	179

This is as close to an even distribution of responses as there was in this test. If we look at the probability curves however we can see that the responses weighted 1 and 2 are never the most probable response at any point in the ability range. There is never a portion of the ability range that these responses can be considered characteristic of. responses 1 and 2 seem to be attracting peolpe that we would expect to choose responses 0 or 3. This fact is confirmed by the large fit statistic for this item. Thus despite the relative equal distribution of responses this item would have to be considered a candidate for revision.

SUMMARY

It is fair to say that this is only the beginning of our attempts to understand the applicability of the Credit model to multiple-choice items. There are many unanswered questions. In particular, is the variance introduced into the estimated ability by using the weighted responses useful in predicting some outcome

such as course success, or are they more highly correlated with some other estimate of this ability such as independent estimates of ability by a panel of experts? Can the amount of this new variance be increased by increasing the number of items like item 12, shown on plot 3, on the test? Currently fewer then 1/3 of the items had probability curves like these. Will the estimation of ability with omits excluded increase the predictive or concurrent validity of the measure? These are certainly substantial questions. The answers to these questions will play an important role in determining the utility of this method of scoring multiple-choice items and the subsequent item analysis. At the current stage of development this is a promising means of dealing with some of the most demanding problems in measurement.

REFERENCES

Bejar, I.I., and Weiss, D.J., A Comparison of Empirical Differential Option Weighting Scoring Procedure as a Function of Inter-Item Correlations. Educational and Psychological Measurement, 1977, 37, 335-340.

Claudy, J.G., Biserial Weights: A New Approach to Test Item Option Weighting. Applied Psychological Measurement, 1978, 2, 25-30.

Coombs, C.H., On the Use of Objective Examinations. Educational and Psychological Measurement, 1953, 13, 308-310.

Coombs, C.H., Milholland, J.E., and Womer, F.B., The Assessment of Partial Knowledge. Educational and Psychological Measurement, 1956, 15, 13-37.

Cronbach, L.J., An Experimental Comparison of the Multiple True-False and Multiple Multiple-Choice Tests. Journal of Educational Psychology, 1941, 32, 533-543.

Cross, L.H., Thayer, N.J., and Frary, R.B., A New method for Administering and Scoring Multiple-choice Tests: Theoretical Considerations and Empirical Results. A paper presented at the annual meeting of the National Council on Measurement in Education, Boston, 1980.

Davis, F.B., Estimation and Use of Scoring Weights for Each Choice in Multiple Choice Test Items. Educational and Psychological Measurement, 1959, 19, 291-298.

Davis, F.B., and Fifer, G., The Effects on Test Reliability and Validity of Scoring Aptitude and Achievement Tests with Weights for every Choice. Educational and Psychological Measurement, 1959, 19, 291-298.

Diamond, J.J., A Preliminary Study of the Reliability and Validity of a Scoring Procedure Based on Confidence and Partial Information. Journal of Educational Measurement, 1975, 12, 129-133.

Dressel, P.L., and Schmid, J., Some Modifications of the Multiple-Choice items. Educational and Psychological Measurement, 1953, 13, 574-595.

Echternacht, G., Reliability and Validity of Item Option Weighting Schemes. Educational and Psychological Measurement, 1976, 36, 301-309.

Guilford, J.P., A Simple Scoring Weight for Test Items and Its Reliability. Psychometrica, 1941, 6, 367-374.

Hambleton, R.K., Roberts, D.M., and Traub, R.E., A Comparison of the Reliability and Validity of Two Methods of Assessing Partial Knowledge on Multiple Choice Tests. Journal of Educational Measurement, 1970, 7, 75-82.

Hendrickson, G.F., The Effects of Differential Option Weighting on Multiple-Choice Objective Tests. Journal of Educational Measurement, 1971, 8, 291-296.

Kelly, T.L., Scoring Alternative Responses With Reference to Some Criterion. Journal of Educational Psychology, 1934, 25, 504-510.

Nedelsky, L., Ability to Aviod Gross Error as a Measure of Achievement. Educational and Psychological Measurement, 1954, 14, 459-472.

Patnaik, D., and Traub, R.E., Differential Weighting by Judged Degree of Correctness. Journal of Educational Measurement, 1973, 10, 281-286.

Poizner, S.B., Nicewander, W.A., and Gettys, C.F., Alternative Responses and Scoring Methods for Multiple Choice Items: An Empirical Study of Probablistic and Ordinal Response Modes. Applied Psychological Measurement, 1973, 2, 83-96.

Ramsey, J.O., A Scoring System for Multile Choice Test Items. The British Journal of Mathematical and Statistical Psychology, 1968, 21, 247-250.

Reilly, R.R., and Jackson, R., Effects of Empirical Option Weighting on Reliability and Validity of an Academic Aptitude Test. Journal of Educational Measurement, 1973, 10, 185-193.

Smith E.R., and Tyler, R.W., Apprising and Recording Student Progress. Harper and Brothers, New York, 1942.

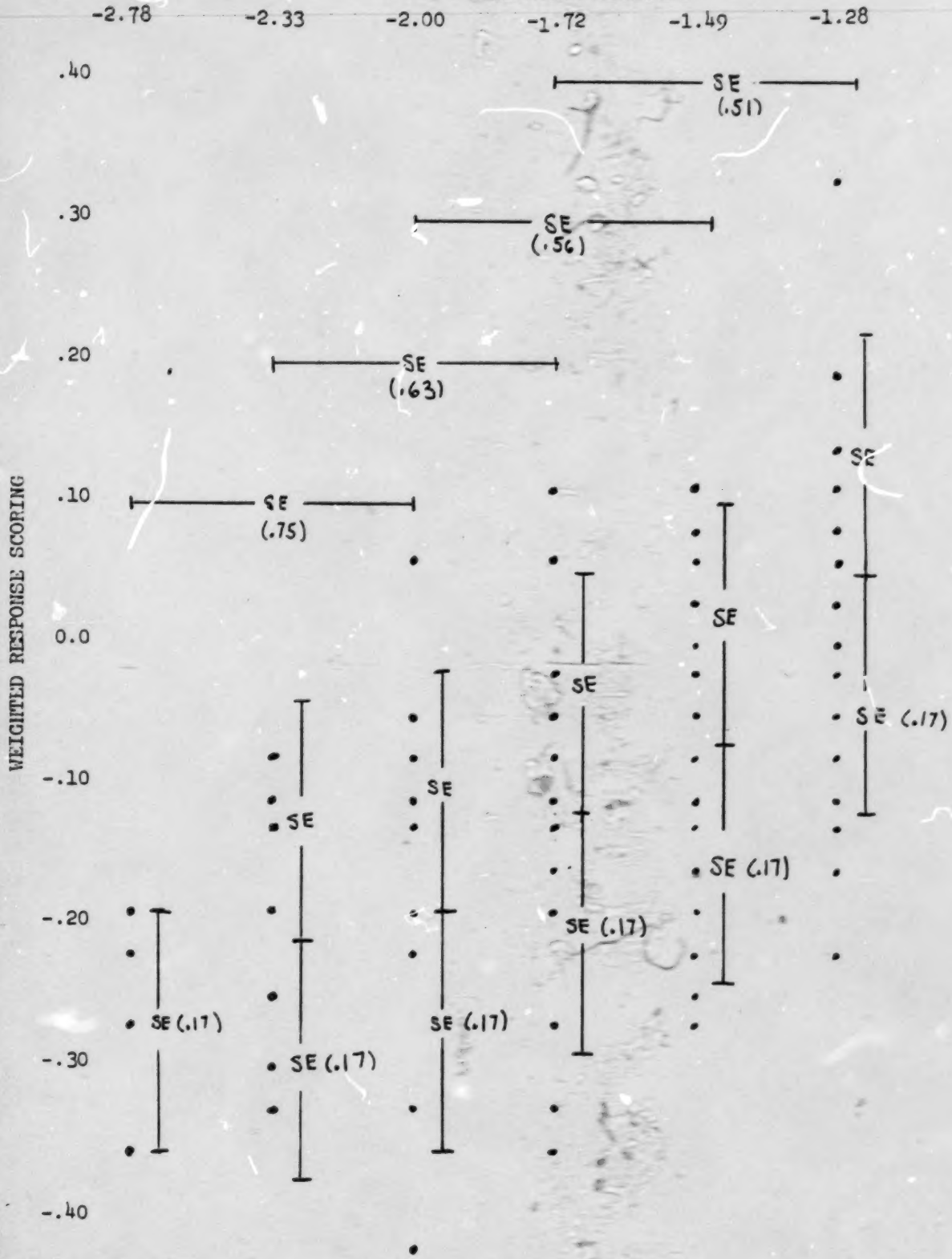
Smith, R.M., A Comparison of A Priori, Empirical, and Dichotomous Scoring Models for Multiple-Choice Items. unpublished manuscript, 1981.

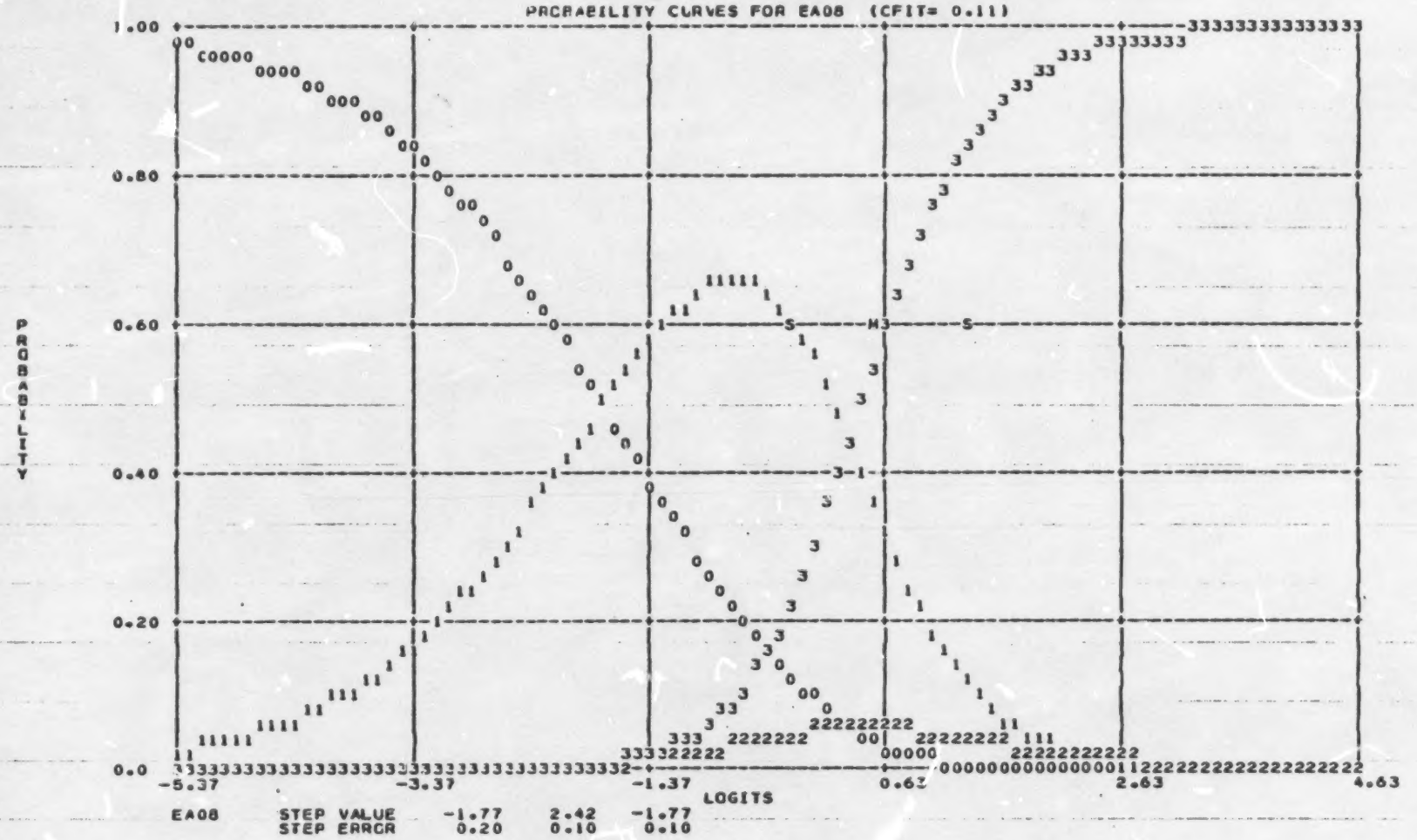
Stalnaker, J.M., Weighting Questions in the Essay-Type Examination. Journal of Educational Psychology, 1938, 29, 481-490.

Thissen, D.M., Information in Wrong Responses to the Raven Progressive Matrices. Journal of Educational Measurement. 1976, 13, 201-213.

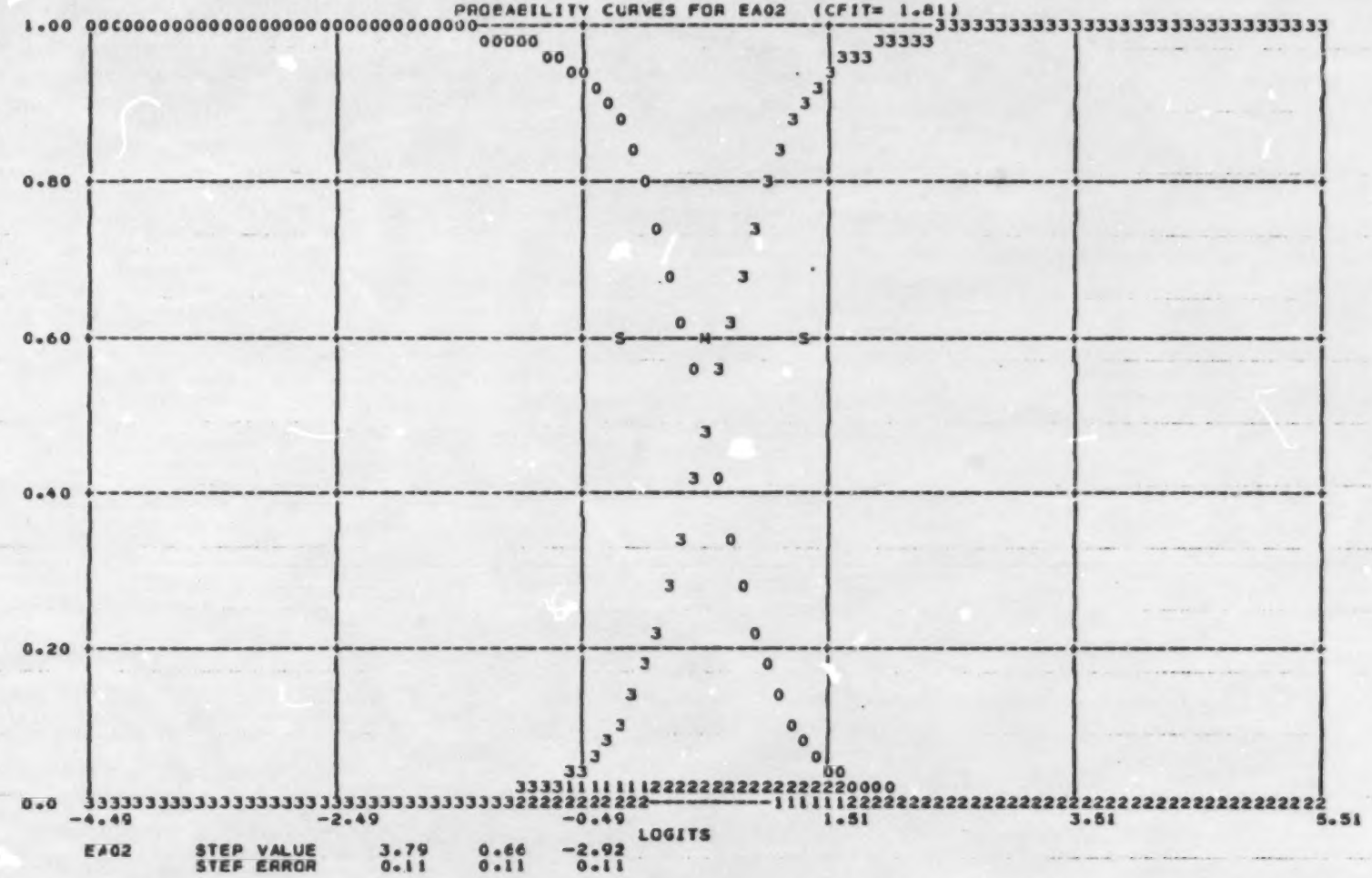
Waters, B.K., The Measurement of Partial Knowledge: A Comparison Between Two Empirical Option Weighting Methods and Rights-Only Scoring. Journal of Educational Research, 1976, , 256-260.

GRAPH 1
ABILITY ESTIMATES DICHOTOMOUS vs. WEIGHTED RESPONSE



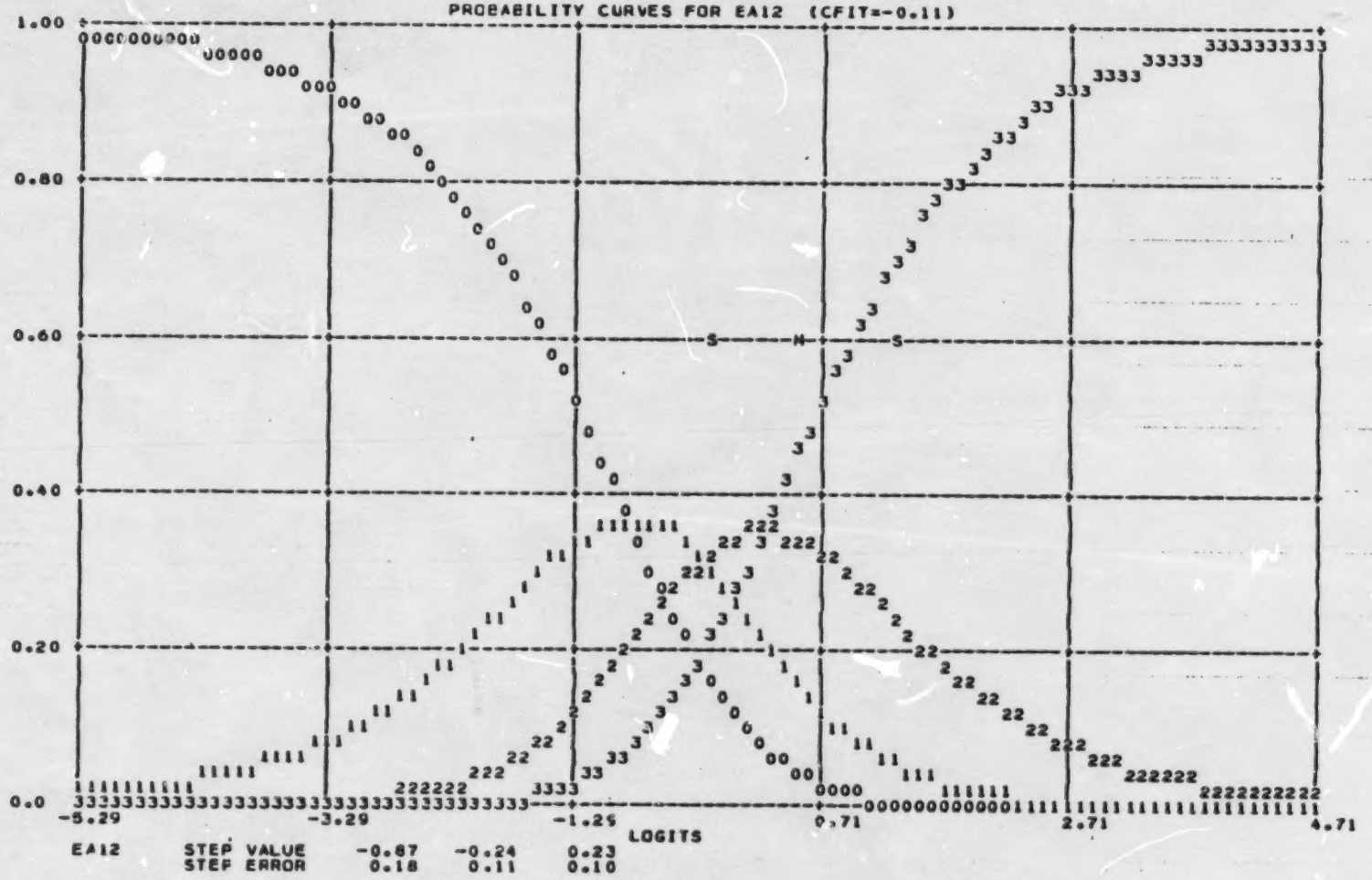

PROBABILITY CURVES FOR EA08 (CFIT= 0.11)
1.00
0.80
0.60
0.40
0.20
0.0
PROBABILITY
-5.37
-3.37
-1.37
0.63
2.63
4.63
LOGITS
EA08 STEP VALUE -1.77 2.42 -1.77
STEP ERROR 0.20 0.10 0.10

PLOT 1

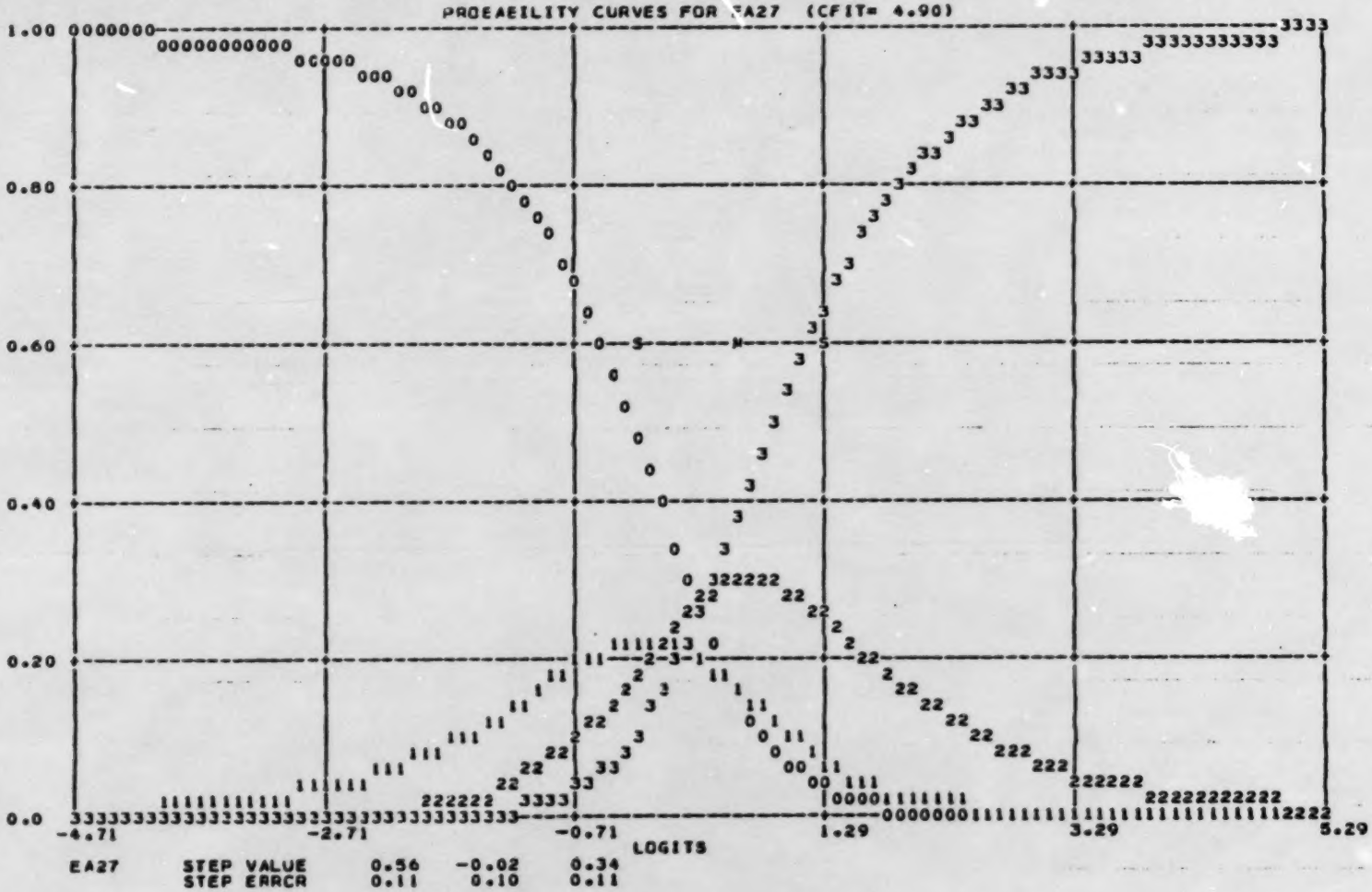


EA02	STEP VALUE	3.79	0.66	-2.92
	STEP ERROR	0.11	0.11	0.11

PLOT 2


PROBABILITY CURVES FOR EA12 (CFIT=-0.11)
PROBABILITY
1.00
0.80
0.60
0.40
0.20
0.0
-5.29
-3.29
-1.25
0.71
2.71
4.71
LOGITS
EA12 STEP VALUE -0.87 -0.24 0.23
STEP ERROR 0.18 0.11 0.10

PLOT 3


PROBEABILITY CURVES FOR EA27 (CFIT= 4.90)
1.00
0.80
0.60
0.40
0.20
0.0
PROBABILITY
-4.71
-2.71
-0.71
1.29
3.29
5.29
LOGITS
EA27 STEP VALUE 0.56 -0.02 0.34
STEP ERRCR 0.11 0.10 0.11

PLOT 4